REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
28-05-2013	Final Report	23-Aug-2012 - 22-May-2013

4. TITLE AND SUBTITLE
Bayesian Kernel Methods for Non-Gaussian Distributions: Binary and Multi-class Classification Problems

5a. CONTRACT NUMBER
W911NF-12-1-0401

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Kash Barker, Theodore B. Trafalis, Cameron A. MacKenzie

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Oklahoma
Board of Regents of the University of Oklahoma
201 David L. Boren Blvd.
Norman, OK 73019 -5715

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
61414-MA-II.3

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Recent advances in data mining have integrated kernel functions with Bayesian probabilistic analysis of Gaussian distributions. These machine learning approaches can incorporate prior information with new data to calculate probabilistic rather than deterministic values for unknown parameters. This paper analyzes extensively a specific Bayesian kernel model that uses a kernel function to calculate a posterior beta distribution that is conjugate to the prior beta distribution. Numerical testing of the beta kernel model on several benchmark data sets reveal that this

15. SUBJECT TERMS
Bayesian, kernel methods, classification, imbalanced data, online learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kash Barker
UU	UU	UU	UU		19b. TELEPHONE NUMBER 405-325-3721

Report Title

Bayesian Kernel Methods for Non-Gaussian Distributions: Binary and Multi-class Classification Problems

ABSTRACT

Recent advances in data mining have integrated kernel functions with Bayesian probabilistic analysis of Gaussian distributions. These machine learning approaches can incorporate prior information with new data to calculate probabilistic rather than deterministic values for unknown parameters. This paper analyzes extensively a specific Bayesian kernel model that uses a kernel function to calculate a posterior beta distribution that is conjugate to the prior beta distribution. Numerical testing of the beta kernel model on several benchmark data sets reveal that this model's accuracy is comparable with those of the support vector machine and relevance vector machine, and the model runs more quickly than the other algorithms. When one class occurs much more frequently than the other class, the beta kernel model often outperforms other strategies to handle imbalanced data sets. If data arrive sequentially over time, the beta kernel model easily and quickly updates the probability distribution, and this model is more accurate than an incremental support vector machine algorithm for online learning when fewer than 50 data points are available.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Baroud, H., K. Barker, C.M. Rocco, and C.A. MacKenzie. 2013. Application of Bayesian Kernel Methods to Network Reliability. To be submitted to Reliability Engineering and System Safety, July 2013.

Baroud, H., K. Barker, C. MacKenzie, and T.B. Trafalis. 2012. Bayesian Kernel Models for Count Data. INFORMS Annual Meeting, Phoenix, AZ, October 2012.

Number of Presentations: 2.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper	
05/28/2013	2.00	Hiba Baroud, Kash Barker, Raychal Lurvey, Cameron MacKenzie. Bayesian Kernel Models for Disruptive Event Data, Industrial and Systems Engineering Research Conference. 2013/05/19 01:00:00, . : ,

TOTAL: **1**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received	Paper

TOTAL:

Number of Manuscripts:

Books

Received	Paper

TOTAL:

Patents Submitted

Patents Awarded

Awards

Best Paper Award, Homeland Security Track for the following:
Baroud, H., K. Barker, R. Lurvey, and C.A. MacKenzie. 2013. Bayesian Kernel Models for Disruptive Event Data.
Proceedings of the 2013 Industrial and Systems Engineering Research Conference, San Juan, PR, May 2013.

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Hiba Baroud	1.00	
FTE Equivalent:	**1.00**	
Total Number:	**1**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Kash Barker	0.11	
Theodore Trafalis	0.11	
Cameron MacKenzie	0.00	
FTE Equivalent:	**0.22**	
Total Number:	**3**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	Discipline
Raychal Lurvey	0.00	Industrial and Systems Engineering
FTE Equivalent:	**0.00**	
Total Number:	**1**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	1.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	1.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):	1.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

Bayesian Kernel Methods for Non-Gaussian Distributions: Binary and Multi-class Classification Problems

Proposal No. 61414-MA-II
Kash Barker (PI) and Theodore B. Trafalis (co-PI), University of Oklahoma
Cameron A. MacKenzie (senior personnel), Naval Postgraduate School

Project Objective

The objective of this project is to develop a Bayesian kernel model built around non-Gaussian prior distributions to address binary and multi-class classification problems.

Contents, in Brief

Approach

Often, classifying observations of data into one class or another (e.g., determining whether an object is a high value target based on pedestrian traffic characteristics) is a difficult problem as observations often cannot be easily distinguished from each other using the basic characteristics exhibited by the observations. Kernel methods provide a means to convert those characteristics into a higher dimensional space that allows for the classes to more readily distinguish themselves. Bayesian methods, in concert with kernel methods, allow for enhanced classification, where the result of a Bayesian kernel model is a *likelihood* (or probability) that the observation falls into a particular class, not simply the class itself. Bayesian methods require a *prior probability distribution* to describe parameters, and observations transform this prior distribution into a more descriptive likelihood of classification. Previous developments in Bayesian kernel models assume a normal distribution as the prior distribution, which can be a problematic assumption for some data sets.

Our approach explored different prior distributions for classification problems: (i) a beta distribution for binary classification problems and (ii) a Dirichlet distribution (an extension of the binomial distribution) for multi-class classification problems, as well as (iii) applying these approaches in online learning environments, where data processing occurs one observation at a time and the classification algorithm improves over time with new observations.

This report essentially summarizes a paper in submission at *Computational Statistics and Data Analysis* [MacKenzie et al. 2013], which is appended to this report.

Scientific Challenges and Opportunities
A primary challenge of this research lies in successfully classifying some difficult classification problems (e.g., one class appears much more frequently than another class) and in improving the performance of the algorithm (across several metrics) relative to existing approaches.

Significance
The significance of this approach, as it turns out, will be to provide similar accuracy to other classification approaches in a significantly reduced amount of computational time, especially for problems of online learning.

Accomplishments
We describe accomplishments with respect to the first three tasks of the proposed work.

Task 1: Binary Classification. We developed the formulation in Eq. (1) to classify the observations of an unknown data point *i* represented by the vector $\mathbf{x}_i$. The probability that data point *i* is positively labeled follows the beta distribution where y_i represents the unknown classification of data point *i* and $\mathbf{y}$ is a vector of *m* known classifications (the training set). Because the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ describes the similarity between two data points, $\mathbf{x}_i$ and $\mathbf{x}_j$, integrating the use of a kernel function and a beta prior distribution improves classification capability. The number of positive and negative data points often differs in a training set, and the probability distribution on $\mathbf{x}_i$ may reflect that the training set has more of one class than the similarity between points as given by the kernel functions. We resolve this problem by adding weighting parameters m_-/m and m_+/m, where m_- and m_+ are the number of negative and positive labels, respectively, in the training set. The parameters $\alpha > 0$ and $\beta > 0$ are the prior distribution parameters for the beta distribution.

$$P(y_i|\mathbf{y}) \sim \text{beta}\left(\alpha + \frac{m_-}{m}\sum_{\{j|y_j=1\}} k(\mathbf{x}_i, \mathbf{x}_j), \beta + \frac{m_+}{m}\sum_{\{j|y_j=-1\}} k(\mathbf{x}_i, \mathbf{x}_j)\right) \quad (1)$$

We tested the beta kernel model on several data sets and compare the results to the relevance vector machine (RVM), the traditional soft-margin SVM [Cristianini and Shawe-Taylor 2000, Shawe-Taylor and Cristianini 2004], and a weighted soft-margin SVM [Chew et al. 2001]. The SVM is a kernel-based linear classifier that uses a relatively small number of vectors to create a boundary between the classes in the feature space. The soft-margin SVM assigns a cost parameter for misclassifications. In the weighted SVM, we assign a different cost for the misclassification of each class: Cm_+/m for the positive class and Cm_-/m for the negative class, where C is a constant cost parameter to

be optimized. We used LIBSVM 3.0 [Chang and Lin 2001] for the SVM models and the code developed by Tipping [2009] for the RVM.

The beta kernel model uses a uniform prior ($\alpha = 1$ and $\beta = 1$) with a weighted likelihood as given in Eq. (1). If the expectation of the posterior probability is greater than 0.5, the unknown point is positively labeled. A non-uniform prior could select α and β so that the expected value of the prior equals the proportion of positively classified data points in the training set, and the threshold could be the expectation of the prior. The non-uniform prior's classifications and the uniform prior's classifications are identical, however, because both classifiers ultimately rely on comparing the summation of the kernel functions of the positively labeled training data points to that of the negatively labeled training data points (the Appendix of MacKenzie et al. [2013] provides a proof of this).

The radial basis kernel function in Eq. (2) was used throughout this work and in MacKenzie et al. [2013], where $\sigma > 0$ is tuned to optimize each classifier. The radial basis kernel is perhaps the most popular kernel function because the image of the function lies on (0,1) and the kernel matrix has full rank [Scholkopf and Smola 2002].

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \tag{2}$$

We applied this new formulation, outlined in MacKenzie et al. [2013], to six data sets available from the University of California-Irvine Machine Learning Repository (Parkinson, Haberman's survival, Arcene, Spam, Transfusion, Breast cancer), one data set from the Princeton University Gene Expression Project (Colon cancer), and one from National Weather Center at the University of Oklahoma (Tornado). We divided each of the data sets into training, tuning, and testing sets. The training set comprises 50 percent of each data set, the tuning set 20 percent, and the testing set 30 percent. In each individual trial, the σ in the kernel function (as well as the cost parameter C in the SVM) is selected that achieves the highest accuracy score in the tuning set. The training and tuning set were combined to retrain the classifier using the optimal σ (and C) and test it on the testing set.

Table 1 displays the mean performance across 200 repetitions for the true positive (TP) rate, the true negative (TN) rate, the accuracy score $\text{Acc} = \sqrt{\text{TP} \times \text{TN}}$, and computational time (in seconds) for the beta kernel approach we developed.

The beta kernel approach had the best accuracy for many of the benchmark data sets. Perhaps more importantly, our approach significantly outperformed the existing approaches on all data sets in terms of computational time.

Task 2. Multi-class Classification. Binary classification problems have two classes (a positive and negative class), but multi-class classification problems have $N > 2$ classes.

We developed the formulation in Eq. (3) for the Dirichlet distribution describing the likelihood of classifying an unknown data point i in one of the N classes, and y_i can be an integer from 1 to N. Given a Dirichlet prior represented by parameters $\alpha_1, \alpha_2, \ldots, \alpha_N$, the weighted kernel approach can be used to derive the posterior distribution that also follows a Dirichlet distribution, where m_{-n} is the number of known data points not in the nth class and m is the total number of known observations. Weighting parameters are included to account for data where the number in each class is not the same.

Table 1. Performance of the binary classification approaches.

Data set	Metric	Beta kernel	RVM	Traditional SVM	Weighted SVM
Parkinson	Acc	0.635	0.068	0.363	0.524
	TP	0.680	1.000	1.000	0.710
	TN	0.710	0.060	0.280	0.620
	Time	0.200	22.880	3.870	4.090
Haberman's survival	Acc	0.487	0.072	0.218	0.450
	TP	0.860	1.000	0.990	0.810
	TN	0.370	0.040	0.120	0.400
	Time	0.280	5.010	20.960	11.010
Arcene	Acc	0.660	0.125	0.125	0.125
	TP	0.760	0.130	0.130	0.130
	TN	0.720	1.000	1.000	1.000
	Time	4.850	38.740	371.540	622.880
Spam	Acc	0.463	0.260	0.351	0.564
	TP	0.330	0.180	0.230	0.480
	TN	0.980	0.990	0.990	0.900
	Time	0.960	348.520	26.1000	27.560
Colon cancer	Acc	0.716	0.250	0.249	0.466
	TP	0.780	0.250	0.250	0.500
	TN	0.830	1.000	1.000	0.900
	Time	0.220	5.690	29.710	29.530
Transfusion	Acc	0.533	0.103	0.108	0.522
	TP	0.470	0.040	0.040	0.430
	TN	0.710	1.000	0.980	0.720
	Time	1.770	66.230	266.010	106.240
Breast cancer	Acc	0.100	0.105	0.105	0.298
	TP	0.110	0.110	0.110	0.510
	TN	0.940	1.000	1.000	0.640
	Time	0.050	46.960	2.840	2.700
Tornado	Acc	0.533	0.103	0.108	0.522
	TP	0.470	0.040	0.040	0.430
	TN	0.710	1.000	0.980	0.720
	Time	5.960	1621.130	115.250	160.690

$$P(y_i|\mathbf{y}) \sim \text{Dir}\left(\alpha_1 + \frac{m_{-1}}{m}\sum_{\{j|y_j=1\}} k(\mathbf{x}_i,\mathbf{x}_j), \ldots, \alpha_N + \frac{m_{-N}}{m}\sum_{\{j|y_j=N\}} k(\mathbf{x}_i,\mathbf{x}_j)\right) \quad (3)$$

We applied this new formulation to four data sets available from the University of California-Irvine Machine Learning Repository (Iris, Wine, Satellite, Steel faults). Table 2 displays the mean performance across 200 repetitions for the overall accuracy score $\text{Acc} = (\prod_{n=1}^{N} \text{Acc}_n)^{1/N}$, where Acc_n is the proportion of observations in the nth class accurately classified. The computational time (in seconds) for the beta kernel approach is also depicted. We developed the beta kernel model with both uniform and non-uniform Dirichlet parameters, and compared it to the multi-class SVM and Classification and Regression Trees (CART) [Hastie et al. 2001].

Our Bayesian kernel approach with the weighted Dirichlet prior distribution did not perform across several metrics as well as the other existing approaches, especially CART. We will continue to explore this approach, but initial results do not appear promising.

Table 2. Performance of the multi-class classification approaches.

Data set	Metric	Beta kernel	RVM	Traditional SVM	Weighted SVM
Iris	Acc	0.939	0.943	0.955	0.945
	Time	0.120	0.110	0.480	0.010
Wine	Acc	0.948	0.949	0.973	0.911
	Time	0.160	0.150	0.700	0.030
Satellite	Acc	0.784	0.834	0.866	0.788
	Time	33.880	33.780	111.680	1.130
Steel faults	Acc	0.733	0.733	0.753	0.693
	Time	7.960	7.800	78.980	1.130

Task 3. Online Learning. Often classification algorithms are deployed in a "batch" setting, where classification parameters are found at once from multiple training cases. This is counter to "online" learning, where processing occurs one observation at a time. Such an approach allows for very large training sets and for updating classification parameters as new data arrive (e.g., sensor data streaming regularly). We demonstrate the application of the beta kernel model to online learning with the benchmark twonorm data set, downloaded from the Delve project at the University of Toronto.

We select two data points for which we assume the characteristics but not the outcomes are known. At each iteration, a unique set of 10 data points whose outcomes are known is used to update beta parameters α and β for each of the two unknown data points. Table 3 depicts the updated α and β and the expected posterior probability.

Figure 1 displays the beta distribution's probability density function as α and β are updated for each of these two data points.

Table 3. Updated parameters for the beta kernel model with twonorm data.

Iteration	Data point 1			Data point 2		
	α	β	$\bar{\theta}_1$	α	β	$\bar{\theta}_2$
Prior	1.00	1.00	0.50	1.00	1.00	0.50
1	1.21	1.35	0.47	1.04	2.32	0.31
2	2.05	1.54	0.57	1.28	3.35	0.28
5	2.18	3.01	0.42	1.31	6.66	0.16
10	4.92	4.97	0.50	1.70	10.47	0.14
20	8.29	8.40	0.50	2.59	19.54	0.12
30	13.50	11.71	0.54	3.59	27.73	0.11

As the classifier receives more information, the first data point is much more likely to result in a positive outcome than the second data point. The expected probability for the first data point is close to 0.5 during all the iterations. Even after 30 iterations, the beta distribution's density function (the dark solid line in Figure 1a) is still wide enough that the probability of a positive classification could be between 0.25 and 0.75. The first data point's expected probability is 0.54. Much uncertainty exists over whether this data point is positively or negatively labeled; however, the posterior probability is much greater than 0.25, the fraction of positively labeled data points in the data set. Updating the parameters for the second data point significantly reduces the uncertainty of this data point's outcome. After only 5 iterations or 50 data points, the expected probability of a positive classification is 0.16. After 30 iterations, the expected probability is only 0.11, and most of the beta distribution's density function is less than 0.25. It seems pretty clear that the second data point is a negative classification.

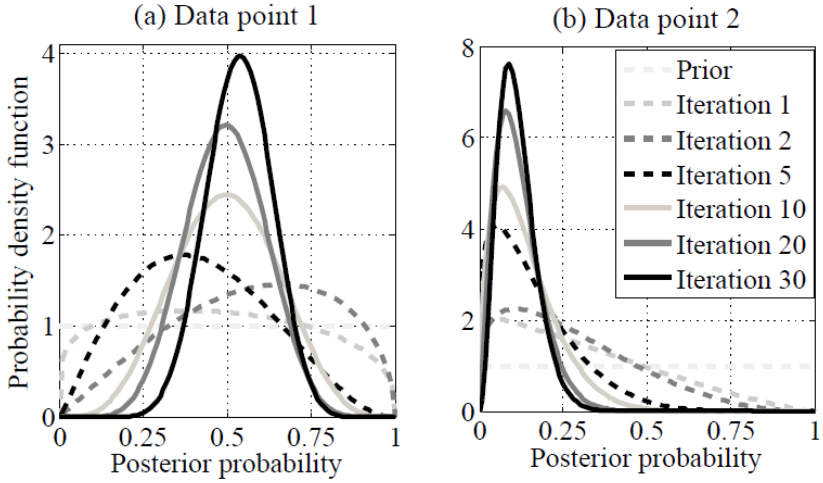

Figure 1. Posterior probability distributions for two different data points in the twonorm data set.

Task 4. Application. We have applied the Bayesian kernel approach to a risk-based problem in network reliability, which could be of interest to the Army. Network reliability problems are typically solved with a max-flow min-cut algorithm following the complete or partial disruption of one or more components in the network [Rocco and Muselli 2007]. That is, an algorithm is performed to determine the connectivity and disrupted flow across the network, where the ratio of disrupted flow to as-planned flow provides a measure of network reliability. Such an algorithm can take a significant amount of computational time, especially for large networks.

We deploy a novel application of our technique to train the Bayesian kernel algorithm with flows along the links of a network (50,000 flows randomly chosen to generate a training set) and use the traditional max-flow min-cut algorithm to determine the connectivity of the network (classified as "desired flow from source to sink node achieved" or "not achieved"). The application of the Bayesian kernel approach can drastically reduce the computational time to determine network connectivity of a disrupted network (or potentially disrupted network). Our very initial results suggest promising results, but more work remains in (i) comparing to the max-flow min-cut algorithm with respect to computational time, (ii) optimizing the tuning parameter (σ) of the radial basis kernel function to improve performance, (iii) optimizing the "decision rule" for classifying an observation based on its probability of falling in that class.

The three measures of performance for the various training and testing sets are provided in Eq. 4. *Acc* is the accuracy of the approach is the proportion of observations that were correctly classified (when an observation with a 0.51 probability or more of the network being in an operating state is classified as such). *Sens* is the quantitative descriptor of sensitivity, and *Spec* is the quantitative descriptor of specificity. *TP*, *TN*, *FP*, and *FN* are the counts of true positive classifications, true negative, false positive, and false negative observations, respectively.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}, \text{Sens} = \frac{TP}{TP + FN}, \text{Spec} = \frac{TN}{TN + FP} \tag{4}$$

Not surprisingly, as the training size increases, predictive accuracy improves, according to Table 4. This represents only one run of training and testing sets, but more runs should be performed to gain a better understanding of average performance.

In a different initial experiment, three different runs were performed with varying training, tuning, and testing set sizes. For the tuning phase, an optimal radial basis function parameter σ was computed based on the maximum value of the accuracy measure. In the testing phase, the tuning and training sets were combined into one training set with the σ found in the tuning phase. Set 1 contains 5% tuning, 5% training, and 90% testing data. Set 2 contains 10%, 20%, and 70% respectively, and Set 3 contains

20%, 30%, and 50% respectively. In both phases the prior parameters were compute such that the mean of the beta distribution is equal to the proportion of the positive classifications.

Results of this initial experiment are found in Table 5. The reduction in testing accuracy is puzzling, and we will explore this further with more experimentation.

Army-specific applications of this idea could include communication and transportation networks or PERT-type project management networks.

Table 4. Performance metrics of one run of different sized training sets at RBF kernel parameter $\sigma = 0.5$ and $\alpha = \beta = 1$, network reliability example.

Training size	Metric	Beta kernel
5000	Acc	0.918
	Sens	0.879
	Spec	0.935
15000	Acc	0.920
	Sens	0.842
	Spec	0.952
25000	Acc	0.924
	Sens	0.869
	Spec	0.947

Table 5. Updated parameters for the beta kernel model with three sets identified by (Tuning/Training/Testing), network reliability example.

Iteration	Set 1 (5%/5%/90%)			Set 2 (10%/20%/70%)			Set 3 (20%/30%/50%)		
	σ	acc	$\frac{\alpha}{\alpha+\beta}$	σ	acc	$\frac{\alpha}{\alpha+\beta}$	σ	acc	$\frac{\alpha}{\alpha+\beta}$
Tuning	0.63	0.94	0.30	0.90	0.96	0.30	0.97	0.97	0.29
Testing	0.63	0.71	0.15	0.90	0.78	0.20	0.97	0.75	0.16

Collaborations and Leveraged Funding

Previously a graduate student when this STIR proposal was originally submitted, Cameron MacKenzie is now an Assistant Professor with the Naval Postgraduate School. The Bayesian kernel approach developed here is a good candidate to address some of the DoD classification problems that Dr. MacKenzie will encounter. One application area of interest to Dr. MacKenzie is adversary identification with specific problems in border security.

Conclusions

This grant explored the usefulness of the beta kernel model and compared the model's accuracy with the RVM (a binary classification algorithm based on Gaussian distributions) and the SVM. The beta kernel model relies on the well-known beta-binomial Bayesian formula, and deploying a kernel function as a measure of similarity between two different data points enables us to apply these updating techniques to classification problems. Incorporating weighting parameters or beginning with a non-uniform prior can help the model correctly classify imbalanced data sets.

The extensive numerical testing of the beta kernel model with the RVM and SVM indicates that the beta kernel model may have some advantages that can be exploited for classification problems. The beta kernel model performs similarly to the SVM and a weighted SVM for the eight data sets in which the minority class composes between 7 and 44% of the data. The beta kernel model consistently performs better than the RVM. If the user desires a probabilistic data mining tool, the beta kernel model may be a superior choice to the RVM. When the minority class comprises only 5% of the data, the beta kernel model generates accuracies on par with those of under-sampling the data combined with either the RVM or SVM. The accuracy of the beta kernel model is significantly better than undersampling and over-sampling, among others, for two of the heavily imbalanced data sets. This suggests that for heavily imbalanced data sets, the beta kernel model should be considered along with under-sampling the RVM or under-sampling the SVM.

The online learning experiment reveals that the beta kernel model outperforms the RVM and LASVM (an incremental learning version of the SVM) if 50 or fewer data points are available. Finally, the beta kernel model calculates posterior probabilities very quickly and runs faster than the RVM and SVM, both of which rely on solving optimization problems.

As this work represents the first extensive analysis and testing of the beta kernel model, we believe the model can potentially become a useful tool in machine learning. The beta kernel model may not provide significant advantages for classifying data sets where the number in each class is relatively the same, but the model carries other advantages, like fast run-times. If the data set is heavily imbalanced, the beta kernel model may be the most accurate. If the data arrive incrementally, the model easily and quickly updates to incorporate the new data and can be relatively accurate with just a few data points.

Unfortunately, multi-class classification with a Dirichlet prior distribution did not produce favorable results, though future work may improve this initially disappointing result.

The novel application of our approach to network reliability as a less computationally expensive alternative to max-flow min-cut algorithms has initially

shown promising results. A paper further elucidating this idea will be finished in the Summer of 2013.

Technology Transfer
None.

Future Work
This initial STIR funding has led to several future research ideas that we hope to explore.

Hierarchical Bayesian Kernel Methods. As with any statistical analysis, the ability of Bayesian methods (whether or not they are integrated with kernel methods) suffers when data describing events of interest are sparse. This is particularly true in the analysis of risk of low-likelihood, potentially high-impact events: very little data exist to describe such events and performing any type of statistical analysis poses challenges. Another extension of Bayesian methods is the hierarchical Bayesian model whose approach would borrow data from similar systems or subsystems in order to evaluate extreme events that usually lack the availability of large datasets necessary to estimate parameters.

We would like to pursue, in a longer term research project, the integration of (i) the Bayesian kernel models resulting from the currently funded research project with (ii) hierarchical Bayesian models. Such hierarchical Bayesian kernel (HBK) methods would result in many benefits where an accurate estimation of risk parameters is improved by the use of kernel functions even though direct data might be unavailable.

Count Data Modeling. With the STIR, we explored Bayesian kernel methods for classification problems. A future, and very unique, area we hope to pursue is the application of the Bayesian kernel methods (and also HBK methods) to *count data* as opposed to classification data. Count data describe the number of occurrences of an event over a given time period (e.g., three earthquakes in one year). In particular, we are interested in describing the likelihood of disruptive events in networks. For example, in critical infrastructure systems, our proposed HBK methods can be used to estimate probabilities of component failure given information on past failures of similar components in other systems or subsystems.

Applications in Resilience. Consider the system state transition in Figure 2, describing the onset and eventual recovery from a disruptive event e_j occurring at time t_e. The PI has explored system, and specifically network, recovery and resilience in several recent

works [Barker et al. 2013, Barker and Baroud 2012, Baroud et al. 2013a,b, Pant et al. 2012, Pant et al. 2013].

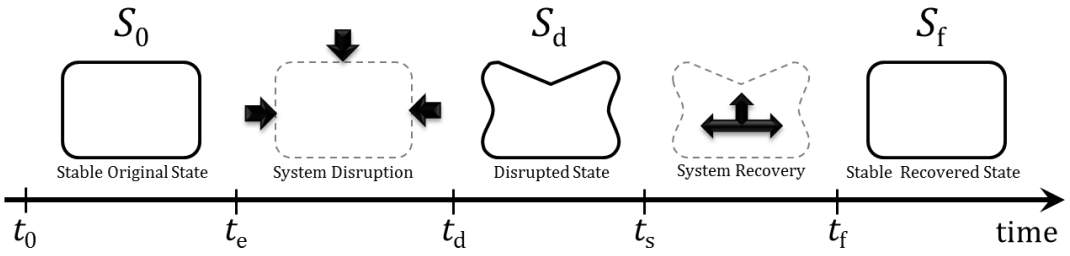

Figure 2. System state transition with time, from original to disrupted to recovered states.

The above works propose, extend, and apply a paradigm for assessing the resilience of a system by quantifying the damage to the network service function $\varphi(\cdot)$. For example, if the network under study is an inland waterway network, $\varphi(\cdot)$ could measure commodity flows across waterway links. Given a particular disruptive event, e^j, Eq. (4) provides a more specific quantification of the value of resilience $Я_F(t_r|e_j)$ evaluated at time $t_r \in (t_d, t_f)$, where set $\mathcal{D}$ is the set of possible disruptive events.

$$Я_{\varphi}(t_r|e_j) = \frac{[\varphi(t_r|e_j) - \varphi(t_d|e_j)]}{[\varphi(t_0) - \varphi(t_d|e_j)]} \quad \forall e_j \in \mathcal{D} \tag{4}$$

In future work, we wish to apply the non-Gaussian Bayesian kernel models discussed resulting from the ARO STIR to model the resilience of disrupted systems. The outcome of Eq. (4), $Я_{\varphi}(t_r|e_j)$, lies between 0 and 1, with 0 representing a completely non-functional system and 1 representing a recovered system. Therefore, a suitable conjugate prior in this case is the Beta distribution, for which the range of the random variable is [0,1]. Eq. (5) is a *conceptual* representation of the Beta probability distribution with parameters $\alpha > 0$ and $\beta > 0$, where Я is the resilience described in Eq. (4) and $B(\alpha, \beta)$ is the beta function. Using a set of covariates that relate to the disrupted system (e.g., system characteristics, the disruption itself, recovery time, cost metrics), the Bayesian kernel model estimates resilience according to the characteristics of each data point.

$$P(Я) = \frac{Я^{\alpha-1}(1-Я)^{\beta-1}}{B(\alpha, \beta)} \tag{5}$$

We feel that system recovery and resilience is an important topic, particularly within the DoD, and this represents an important extension of the Bayesian kernel and HBK approaches.

Products of Research Funding

Papers submitted

MacKenzie, C.A., T.B. Trafalis, and K. Barker. 2013. Non-Gaussian Bayesian Kernel Methods for Binary Classification and Online Learning Problems. In review at *Computational Statistics and Data Analysis*.

Papers in progress

Baroud, H., K. Barker, and C.A. MacKenzie. 2013. Bayesian Kernel Models for Count Data. To be submitted to *Computational Statistics and Data Analysis*, August 2013.

Baroud, H., K. Barker, C.M. Rocco, and C.A. MacKenzie. 2013. Application of Bayesian Kernel Methods to Network Reliability. To be submitted to *Reliability Engineering and System Safety*, July 2013.

Conference papers and presentations

Baroud, H., K. Barker, R. Lurvey, and C.A. MacKenzie. 2013. Bayesian Kernel Models for Disruptive Event Data. *Proceedings of the 2013 Industrial and Systems Engineering Research Conference*, San Juan, PR, May 2013. Best Paper Award, Homeland Security Track.

Baroud, H., K. Barker, C. MacKenzie, and T.B. Trafalis. 2012. Bayesian Kernel Models for Count Data. INFORMS Annual Meeting, Phoenix, AZ, October 2012.

Ph.D. student support (on-going)

Hiba Baroud

References

Barker, K., J.E. Ramirez-Marquez, and C.M. Rocco. 2013. Resilience-Based Network Component Importance Measures. To appear in *Reliability Engineering and System Safety*. < http://dx.doi.org/10.1016/j.ress.2013.03.012>

Barker, K. and H. Baroud. 2012. Proportional Hazards Models of Infrastructure System Recovery. In revision in *IEEE Transactions on Systems, Man, and Cybernetics Part A*.

Baroud, H., J.E. Ramirez-Marquez, K. Barker, and C.M. Rocco. 2013. Measuring and Planning for Stochastic Network Resilience: Application to Waterway Commodity Flows. Submitted to *Risk Analysis*.

Baroud, H., K. Barker, J.E. Ramirez-Marquez, and C.M. Rocco. 2013. Importance Measures for Inland Waterway Network Resilience. Submitted to *Transportation Research Part E*.

Chang, C.-C., and C.-J. Lin. 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chew, H.-G., D.J. Crisp, R.E. Bogner, and C.-C. Lim. 2000. Target Detection in Radar Imagery using Support Vector Machines with Training Size Biasing. *In Proceedings of the 6th international conference on control, automation, robotics, and vision*, Singapore.

Cristianini, N. and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Pant, R., K. Barker, J.E. Ramirez-Marquez, and C.M. Rocco S. 2012. Stochastic Measures of Resilience and their Application to Container Terminals. Submitted to *Annals of Operations Research*.

Pant, R., K. Barker, and C.W. Zobel. 2013. Resilience Estimation Metrics for Interdependent Infrastructure and Industry Sectors. In revision in *Reliability Engineering and System Safety*.

Rocco, C.M. and M. Muselli. 2007. Network Reliability Assessment through Empirical Models using a Machine Learning Approach. *Computational Intelligence in Reliability Engineering: New Metaheuristics, Neural and Fuzzy Techniques in Reliability*. G. Levitin (ed.). Berlin: Springer-Verlag. pp. 145-174.

Scholkopf, B. and A.J. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press.

Shawe-Taylor, J. and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Tipping, M. E. 2001. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, **1**: 211-244.